Title: Development and validation of the EUROFORGEN *NAME* (North African and Middle Eastern) ancestry panel

Authors: V. Pereira, A. Freire-Aradas, D. Ballard, C. Børsting, V. Diez, P. Pruszkowska-Przybylska, J. Ribeiro, N.M. Achakzai, A. Aliferi, O. Bulbul, M.D. Perez Carceles, S. Triki-Fendri, A. Rebai, D. Syndercombe Court, N. Morling, M.V. Lareu, Á. Carracedo, The EUROFORGEN-NoE Consortium, C. Phillips

PII: S1872-4973(19)30161-9
DOI: https://doi.org/10.1016/j.fsigen.2019.06.010
Reference: FSIGEN 2109

To appear in: *Forensic Science International: Genetics*

Received date: 8 April 2019
Revised date: 7 June 2019
Accepted date: 13 June 2019

Please cite this article as: { https://doi.org/

Development and validation of the EUROFORGEN *NAME* (North African and Middle Eastern) ancestry panel

V. Pereira [a,1], A. Freire-Aradas [b,1], D. Ballard[c], C. Børsting[a], V. Diez[a], P. Pruszkowska-Przybylska[a,h], J. Ribeiro[a], N.M. Achakzai[b], A. Aliferi[c], O. Bulbul[d], M.D. Perez Carceles[e], S. Triki-Fendri[f], A. Rebai [f], D. Syndercombe Court[c], N. Morling[a], M.V. Lareu [b], Á. Carracedo[b,g]; The EUROFORGEN-NoE Consortium; C. Phillips[b*]

[a] *Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Frederik V's Vej 11, DK-2100 Copenhagen, Denmark*
[b] *Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain*
[c] *Faculty of Life Sciences and Medicine, King's College, London, UK*
[d] *Institute of Forensic Science, Istanbul University, Istanbul, Turkey*
[e] *Department of Legal Medicine, University of Murcia, Murcia, Spain*
[f] *Centre of Biotechnology of Sfax, Bioinformatics Research Group, Sfax, Tunisia.*
[g] *Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia*
[h] *Department of Anthropology, Faculty of Biology and Environmental Protection, University of Łódź, Poland*

[*]Corresponding author： E-mail address: c.phillips@mac.com (C. Phillips).

[1]These authors contributed equally to the study.

Highlights

- Custom-built ancestry panel of 111 AIM-SNPs developed for North African and Middle Eastern populations ('NAME' panel)
- Independent evaluation and validation in three European member laboratories of the EUROFORGEN consortium
- A maximum level of differentiation of six major geographical regions is obtained when using 237 AIM-SNPs (NAME panel and Global AIMs panel combined)

Abstract

Inference of biogeographic origin is an important factor in clinical, population and forensic genetics. The information provided by AIMs (Ancestry Informative Markers) can allow the differentiation of major continental population groups, and several AIM panels have been developed for this purpose. However, from these major population groups, Eurasia covers a wide area between two continents that is difficult to differentiate genetically. These populations display a gradual genetic cline from West Europe to South Asia in terms of allele frequency distribution. Although differences have been reported between Europe and South Asia, Middle East populations continue to be a target of further investigations due to the lack of genetic variability, therefore hampering their genetic differentiation from neighboring populations. In the present study, a custom-built ancestry panel was developed to analyze North African and Middle Eastern populations, designated the 'NAME' panel. The NAME panel contains 111 SNPs that have patterns of allele frequency differentiation that can distinguish individuals originating in North Africa and the Middle East when combined with a previous set of 126 Global AIM-SNPs.

Keywords: biogeographic ancestry; AIMs; SNPs; MPS; Middle Eastern populations

1. Introduction

The Middle East occupies a central geographic location between the populous regions of Africa, Europe, Central and South Asia; and is characterized by a wide range of climates and landscapes. Agriculture was first developed around 10,000 years ago in the Fertile Crescent and Nile Valley, and subsequently spread across Europe and Asia [1]. Since then, the area has been subject to different population migrations connecting Europe, Asia and Africa. Egyptian, Sumerian, Babylonian, Phoenician, Persian, Greek, Roman, and Ottoman Empires all had major settlements or originated in the region [1,2]. As a consequence, the demographic and genetic structure of Middle Eastern populations has changed dramatically over time. Studies of haploid markers have revealed a sex-biased mosaic of diversity in the area [3]. Y-chromosome analyses indicate strong North African and East Mediterranean components, while mtDNA variation shows significant proportions of European lineages [4]. Studies suggest that geography was not the only cause of genetic variation, and that patterns of population substructure were influenced by factors such as culture and religion. The predominantly Muslim Middle Eastern populations show signs of genetic admixture with African populations, while Christian groups have higher proportions of Western European ancestry [5,6]. As a consequence, current political borders do not properly reflect the underlying distribution of genetic structure or ancestry in the region (Fig. 1). Therefore, the Middle East is an area of considerable interest from the population genetic point of view, yet genetic information is lacking for many of the region's populations.

Knowledge of individual ancestry can be an important factor in genetic studies. In clinical genetics, analysis of ancestry can detect population structure among case and control samples that may confound variation associated with disease susceptibility. In population genetics, estimating ancestry is a key step in the inference of the geographical origin of individuals and may reveal levels of genetic admixture in populations or the dynamics of recent human migration. In forensic genetics, ancestry analysis can serve as an additional tool in crime investigations. In situations with few or no investigative leads, genotyping of ancestry-informative markers (AIMs) can provide information on the biogeographic ancestry of the donors of trace samples from crime scenes [7–9]. AIMs are predominantly binary SNPs that show highly differentiated allele frequencies among population groups in different geographic regions [7]. The forensic AIM panels published to date may predict the ancestry of an individual based on one of five continental origins [8–11]. In recent years, there

are few studies to distinguish the geographically intermediate regions (such as Middle East, Eurasia, South Asia) from other major continental populations [12–16]. However, differences in genetic variation between closely located populations are small, especially when a long history of gene flow has existed between them. Therefore, fine-scale population differentiation will require the analysis of a greater number of AIMs; of which many are likely to be found in recent, more broadly-based SNP variation surveys of the same geographic regions [7–9;12 –15].

The European Forensic Genetics Network of Excellence (EUROFORGEN-NoE) has developed a panel of ancestry markers specifically designed to help distinguish populations of the Middle East from those of other regions including the geographically closely-sited population groups of Europe and South Asia. In this study, the term 'Middle East' was applied to populations in the regions defined by the area outlined in Fig. 1; extended when possible westwards from Egypt to include North African regions bordering the Mediterranean with the populous countries of Libya, Tunisia, Algeria and Morocco. The term Middle East can often be geographically inconsistent and lacking clear demarcation from the barriers to large-scale population movements that define the main continental regions, such as oceans and mountain ranges. Except for the Sahara, there are only weak barriers to mass movement within the regions shown in Fig. 1, but for convenience these were defined in the current study as: northwest by the Bosphorus straits; east by the Hindu Kush, and north by the sparsely populated Eurasian Steppe. The relationship between North Africa and the Middle East regions east of the Nile is also complex, but because we selected and combined AIMs informative for each area, study populations come from both ends of this range.

We describe the development of a forensic multiplex that has been designated the '*NAME'* panel (i.e. North African-Middle East informative). The NAME panel contains 111 SNPs showing patterns of allele frequency differentiation that can differentiate individuals originating in North Africa and the Middle East from other population groups when combined with a previous set of 126 Global AIM-SNPs [11]. Several recent studies have also discovered Middle East informative AIMs [12–15], but in this study a supervised screen of 650,000 SNP array loci, rather than a literature search, was used to identify suitable candidates.

2. Materials and methods

2.1. Samples, DNA extraction and quantification

Study samples comprised 725 donors originating from eighteen populations that span the broadly positioned NW European-SE South Asian axis (i.e. a line running across Eurasia characterized by a lack of geographic barriers to population movement), plus 29 Algerians from the HGDP-CEPH human genome diversity panel, a widely used set of 944 cell line DNAs from worldwide sampling of 52 populations [17]. This axis has revealed an allele frequency gradient in AIMs identified in previous studies with a degree of discontinuity (i.e. a steepened frequency cline) between Iran/Iraq and Afghanistan [16]. However, it should be noted that this is not a consistently observed phenomenon and such a distribution of variation may be dictated in part by the allele frequency differences observed in loci at the extremes of the range (e.g. marked differences in a SNP allele frequency between European vs. East Asian populations are likely to show a gradient, whereas similar frequencies will not). Population details are compiled in Supplementary Table S1. The samples were collected from the EUROFORGEN laboratory biobanks of: i) Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen (UCPH); ii) Forensic Genetics Unit, University of Santiago de Compostela (USC); and iii) King's Forensics, King's College London (KCL). The work was approved by the Danish ethical committee (H-1-2011-081 and H-3-2012-023) and the BDM Research Ethics Subcommittee (HR-15/16-2989) (KCL).

Samples were extracted from 200 µL of whole blood using the DNA Blood Mini Kit (Qiagen, Hilden, Germany) as recommended by the manufacturer (UCPH); using classical extraction phenol/chloroform from blood stains (USC) or using the EZI DNA Investigator Kit (Qiagen) for buccal swab extraction (KCL). DNA concentrations ranged from 10 to 200 ng/µL.

Additional online SNP genotype data was accessed from the 1000 Genomes Project (N=2,233, excluding admixed population samples) and the HGDP-CEPH diversity panel (N=699). See details in Supplementary Table S1.

2.2. AIM-SNPs selection

Selection of AIM-SNPs for the NAME panel was based on screening of HGDP-CEPH SNP genotypes in relevant population samples for highly differentiated genotype distributions, assessed by calculating Rosenberg's informativeness for assignment metric: I_n using in-house scripts and ranking the most divergent SNPs (δ, and Fst values were also calculated and provide comparable measures of population differentiation [18,19]). The I_n metric was calculated for HGDP-CEPH Algerians vs combined European, South Asian, East Asian or African population data. Separately, the three HGDP-CEPH Israeli populations (Israeli Arab: Bedouin; Druze; and Palestinian samples) were compared to European, South Asian, East Asian and African populations to identify the most divergent SNPs for this part of the Middle East. Compiled candidate AIMs comprised the 5% most divergent SNPs from each chromosome for each comparison and these were 'balanced' as far as possible (i.e. the ratios of African-divergent, European-divergent, etc. were equilibrated to reduce bias towards a single group comparison). When SNP clusters were commonly found in a gene or region, the most informative AIM was selected (minimum 1 Mb spans between loci). North African comparisons were more divergent than Middle East comparisons, and an approximately 40:60 ratio of AIMs (for those population groups respectively) was established to maximize the informativeness of the final NAME panel as a whole. Variant data came primarily from the genotyping studies of Li et al. [20] that analyzed ~650,000 SNPs, accessed using the SPSmart SNP browser ([21], querying the "CEPH Stanford HGDP" database). Although additional variant data relevant to the Middle East regions is now available, scrutiny of 650,000 SNPs genotyped in four populations provides a good basis for identifying the most informative AIMs, since such a large set of SNPs is likely to comprehensively cover all of the most divergent genomic regions between these populations and others.

SNPs that showed very high levels of divergence between Africa and Europe, or Europe and South Asia had been previously identified and selected for the Global AIMs panel for continental differentiations [10,11]; so were excluded from the compiled lists. Alternative SNPs with near-identical allele frequency distributions as a result of LD-block correlations were previously identified [22]. These loci can provide substitute candidates in the case of problems with SNP multiplexing and this strategy was used to find new AIMs when necessary.

2.3. AIM-SNPs amplification and genotyping with MassARRAY®

A total of 111 AIM-SNPs distributed in four multiplexes (ME1-ME4) were amplified by UCPH and USC using the iPLEX® Gold Kit (Agena Bioscience GmbH, Hamburg, Germany) in a final reaction volume of 5 µL. Information about the markers and primers included in each iPLEX multiplex is outlined in Supplementary Table S2.

The PCR amplification mix consisted of 1 µL DNA, 0.5 µL 10x Buffer, 0.1 µL dNTP mix (25 mM), 1.3 µL primer mix (0.5 µM each), 0.2 µL HotStarTaq (5 U/µL) and two different final concentrations of $MgCl_2$ depending on the multiplex (3 mM for ME2, ME3 and 4 mM for ME1, ME4). The following thermal cycling conditions were used: denaturation at 94˚C for 2 min followed by 45 cycles of 20 s at 94˚C; 30 s at 56ºC (for ME1 and ME2) or 62ºC (for ME3 and ME4), 1 min at 72˚C, and a final extension of 3 min at 72˚C.

The PCR products were treated with Shrimp Alkaline Phosphatase (SAP) to dephosphorylate the remaining dNTPs from the PCR. Each reaction contained 0.17 µL SAP buffer, 0.30 µL SAP enzyme and 1.53 µL of water. The SAP reaction was carried out at 37ºC for 40 min and 85ºC for 5 min.

The SBE reaction was carried out with 7 µL SAP-treated PCR products and 2 µl iPLEX® mix (Agena Bioscience). The iPLEX® mix contained 0.2 µL 10x iPLEX® buffer, 0.2 µL iPLEX®-Termination mix, 0.94 µL primer mix (DNA Technology, Denmark), 0.04 µL iPLEX®-enzyme and 0.62 µL water. The SBE reaction was carried out with the following conditions: denaturation for 30s at 94˚C followed by 40 cycles consisting of 3 steps: 5s at 94˚C, 5s at 52˚C and 5s at 80˚C, where steps 2 and 3 were repeated 5 times in each cycle. The final extension consisted of 3 min at 72˚C. Samples were analyzed with the MassARRAY® System (Agena Bioscience) using the autorun settings. All samples were run in duplicate.

2.4. Data analyses

The genotype calls were obtained by Typer 4.0.20 software (Agena Bioscience). Genotype calls were analyzed with R statistical software (R core team, version 2.13.0) using the following parameters: signal-to-noise ratio (SNR) >5, allele peak height >1, allele balance (allele balance= (height allele1-height allele2)/(height allele1+height allele2) >|0.2| and >|0.8| for

heterozygotes and homozygotes, respectively [23]. All genotypes were compared between runs and consensus profiles were generated.

2.5. Statistical analysis

Allele frequencies and deviations from Hardy–Weinberg expectations (HWE) for all markers and for each of the 19 populations, were calculated in the Arlequin v.3.5 software [24] using the Exact test and performing 1,000,000 Markov chain steps. Correction for multiple testing was adjusted according to Bonferroni [25]. Principal component analyses (PCA) were performed using a custom script written in R 3.3.1 (http://www.R-project.org/). The genetic ancestry was inferred using the software STRUCTURE v.2.3.4 [26,27]. Analyses were carried out using 100,000 steps of burn-in followed by 100,000 MCMC steps. Six clustering models were considered (K=2-7). We applied an admixed model with correlated allele frequencies and prior labeling of reference populations (i.e. applying the setting: POFLAG = 1) while treating study samples as unknown (POPFLAG = 0). Ten iterations per cluster-model were tested. Plots were constructed using CLUMPAK [28]. Cross-validation was used to estimate the classification success of marker sets by removing each SNP profile and classifying it with the remaining reference data in *Snipper* (http://mathgene.usc.es/snipper/analysispopfile2_new.html). The *Snipper* web portal performs a naïve Bayes analysis that produces a likelihood ratio of the two lowest probabilities based on the principle that SNP allele frequencies in any one population can be directly equated to the probability of origin from that population, when the alleles are present in the profile.

3. Results

3.1. Genotyping performance

Allele frequencies for the 111 SNPs in the nineteen study populations are listed in Supplementary Table S3. HWE was assessed and after Bonferroni correction (p-value: $9{,}00901E^{-05}$) three SNPs (rs7873963, rs896401 and rs10862511) were removed from subsequent analysis for those populations detected to be out of equilibrium (converted to 'NA' values in the corresponding genotype table). The SNP rs7873963 largely failed amplification. In the case of rs896401 and rs10862511, both markers had an excess of homozygotes. Poor quality results were commonly observed for these three SNPs using the MassARRAY. The SNP rs896401 was associated with

another SNP (rs551775849) positioned 17 bp 5′ to the rs896401 locus. Variants in rs551775849 were only found in association with the rs896401-C allele, and since the rs896401 SBE primer included rs551775849, this may explain the low number of observed heterozygotes.

3.2. Patterns of population divergence

Population divergence values were estimated using 2D PCA plots (Fig. 2A/B) for the study samples (N=754) compared to five reference groups: Africans (AFR, N=210), Europeans (EUR, N=191), Middle East populations (ME, N=134), South Asians (SAS, N=270) and East Asians (EAS, N=307). Reference groups included samples either from 1000 Genomes (1K) and HGDP-CEPH for all the clusters, or in the case of ME, only individuals from the CEPH panel, due to the absence of coverage for this region in 1000 Genomes. Corresponding three-dimensional PCAs can be found in Supplementary Fig. S1 for the nineteen study populations. PC1, PC2, and PC3 describe 22.48%, 8.03%, and 3.99% of the variation, respectively. The AFR, EUR and EAS clusters are clearly separated, while ME and SAS samples cluster in the middle of these three groups, and although some overlap is present, two different clusters for these population groups are discernible. Turkish samples were clearly separated from the EUR cluster, showing closer proximity to ME. In the North African study populations, data points for Morocco and Libya match the ME cluster. However, Algeria shows a different pattern, with greater proximity to Africans discernible. Somalis in Fig. 2B are positioned between AFR and ME; while Greenlanders display a broader distribution between EUR and EAS clusters. Roma were included in our study with a small sample size (N=16) and representing individuals derived from different geographical locations (Bulgaria, Spain, Romania and former Yugoslavia). Supplementary Fig. S2 shows their PCA plot individually, and the absence of consistent patterns observed in this plot could be explained by these varied geographic origins.

3.3. Analysis of genetic structure

In order to assess the genetic structure of study samples, STRUCTURE analyses were performed based on the previous reference groups. The same five reference clusters were used to evaluate the nineteen study populations. Fig. 3 shows the STRUCTURE plots for K=2 to K=4. At K=2, AFR co-ancestry was detected in Somalia as the highest membership proportion (average: 0.7387), while ME and SAS displayed some AFR cluster membership components, especially high

in Algerians (average: 0.6819). When applying a three-population model (K=3), Greenlanders display the highest EAS co-ancestry (average: 0.4949). At K=4, a fourth cluster defines the reference ME and SAS populations, but only forms the major membership (0.6447 and 0.4283 respectively). Study ME and SAS populations of Azerbaijani, Iraqi, Kuwaiti, SE Arabian Peninsula, Afghanistani, British Pakistani, Indian and Roma display similar patterns.

To improve the detection of the ME and SAS clusters, a previous set of AIM-SNPs (126 Global AIMs) [11] was analyzed together with the 111 NAME SNPs. Fig. 4 shows the corresponding STRUCTURE plots for K=4 to 6. When assessing K=4, a new uniform cluster appears in comparison to analyses of the previous 111 SNPs, corresponding to the SAS cluster, maintained as a single cluster in all subsequent K-cluster models. For K=5, two minor cluster plots are obtained after running CLUMPAK, one of them (5A) depicting America as a new cluster, and a second (5B), for ME, which appears as a consistent population group. These patterns become more stable at K=6 providing resolution of the six groups: AFR, EUR, ME, SAS, EAS and AME. In these analyses, North AFR can be observed as an independent group showing co-ancestry between AFR and ME. Lastly, Oceanians (OCE) display a mixed pattern between SAS and EAS, and did not form similar distinct clusters compared to either group.

3.4. Bayesian classification analysis

The *Snipper* forensic classifier online tool was used to cross-validate individuals from the six main continental populations that had displayed consistent clusters in STRUCTURE analysis, applying genotype data from a combined set of 237 AIM-SNPs (111 NAME loci plus 126 Global AIMs). Individuals from each population group were correctly classified in full or in high proportions; in descending order of assignment success: AFR (100%), ME (100%), EAS (99.80%), SAS (99.39%), EUR (93.64%) and AME (90.59%). Therefore, using two panels of AIMs almost all African, Middle East, East Asian and South Asian individuals can be distinguished, but Europeans and Native Americans have levels of incorrect assignment of 7-10%, due to the addition of Middle East as a population of origin compared to the closely related origins of Europe and South Asia. When cross-validating individuals with this SNP set using seven populations (adding North Africa, NAFR), classification success shows the same effect of reducing, this time for ME, the most closely related population group to North Africa, with assignment success rates of: AFR (100%), NAFR (100%), EAS (99.80%), SAS (99.39%), EUR (93.64%), AME (90.59%) and ME (77.61%). From a forensic

casework point-of-view, careful interpretation of the STRUCTURE results obtained from unknown DNA donors would need to be made, alongside Bayesian classification approaches such as *Snipper*; and taking account of the value of the likelihoods obtained and the choice of possible populations of origin. Despite such interpretive complexities, the cross validation results suggest that between 90-100% of individuals would be correctly assigned when testing six possible populations of origin in *Snipper* (i.e. not distinguishing between NAFR and ME). The patterns in PCA and STRUCTURE analyses described in sections 3.2 and 3.3, suggest NAFR and ME individuals would be differentiated in many cases, when using an enlarged set of AIMs, and that individuals from South Asia can be successfully distinguished in almost all cases.

4. Discussion

The differentiation of the Eurasian sub-population groups of Europe, North Africa, the Middle East and South Asia represents the single most common operational request made by crime investigators for information on the likely geographic origin of unidentified suspects. However, in terms of population divergence and a shared recent genetic history, these sub-groups are the most closely related and least differentiated of any world region of similar geographic scale. These patterns of minimal divergence within Eurasia reflect the high degree of population movement described in the introduction, resulting from absence of geographic barriers to migration and trade. The distance by sea between the Indian sub-continent and Middle East regions is not much further than the distance across the Mediterranean, so population admixture has been a constant force shaping modern patterns of genetic variability across Eurasian regions. With such factors influencing the populations we studied, the compilation of 111 SNPs into a multiplex complementary to the original Global AIMs panel has not brought the degree of differentiation which is capable of providing clear-cut, unequivocal ancestry assignments that will be informative for investigators. Nevertheless, our studies indicate that Middle East and North African populations can be better differentiated with dedicated SNP panels, than from data obtained with a large-scale ancestry panel dedicated to the major continental population groups alone.

The presence of a finely graded and uniform allele frequency cline running from the northwest of Europe, through the Middle East towards the southern half of South Asia, has been a major factor in the difficulty of differentiating populations at different points on the cline. When previously developing the *Eurasiaplex* forensic AIM SNP multiplex to distinguish European and

South Asian populations [16], the position of each sub-group at opposite extremes of the allele frequency cline that was detected, helped to maximize the statistical inference of geographic origin in a wide range of study populations from these regions, but the *Eurasiaplex* panel could not differentiate Middle East populations from either Europe or South Asia with sufficient statistical power. It is significant that despite increasing AIM numbers almost five-fold, from 23 SNPs in *Eurasiaplex* to the 111 reported here, Middle East populations were not always differentiated from other Eurasian populations when applying PCA, STRUCTURE or Bayes analyses. It is important to note that such clines are not guaranteed by the geography or a generally observed pattern. Analysis of the HGDP-CEPH Middle East populations using very large arrays of 650,000 SNPs also finds mixed genetic cluster patterns in STRUCTURE and a degree of overlap amongst points in PCA-type plots for neighboring Eurasian populations. In the 2008 study of Li et al. [20], Middle East populations showed joint cluster memberships at K=7 between a seventh Middle East inferred cluster, but also European cluster membership proportions at similar or higher levels, and reduced but detectable South Asian cluster membership proportions (Fig. 1 of [20]). Similarly, the PCA-type patterns of the same samples analyzed with the identical 650,000 SNP array data, using multi-dimensional scaling (MDS) performed by Kayser and de Knijff in 2011 [Fig. 2a, 29], indicates a broadly spaced spread of points for Middle East samples that partially overlap with Europe and are positioned close to Africa for a small set of Algerian samples. In a recent study, 86 AIM SNPs were compiled to differentiate Southwest Asia and Mediterranean populations from Eurasian and African populations [15]. The study results show that the Middle Eastern and North African populations tend to cluster close to the Southwest Asia and Southern European populations (Fig. 4C in [15]). Therefore, the difficulties we and others have encountered to adequately differentiate Middle East and North African populations from those of other regions in Eurasia, stem from their population characteristics, not from an insufficient number of markers in a multiplex designed to work with forensic DNA levels.

As forensic SNP analysis moves towards the larger multiplexes possible with massively parallel sequencing (MPS) technologies, this will shape strategies to develop a more generally useful ancestry panel, appropriate for the population characteristics of many urban areas in Europe and North America, where individuals from many different regions of Eurasia form a large proportion of the demographic profile. It is likely to be the case in practice that a significant number of individuals will not be differentiated with sufficient statistical certainty to gain a secure inference of ancestry within the western regions of Eurasia of Europe, North Africa and the Middle East. It could also be viable to adopt a likelihood threshold approach to ensure any inferences made are based on the highest

likelihoods. This method was successfully adopted for the *Eurasiaplex* SNPs and only a small proportion of individuals from South Asia gave insufficiently informative likelihood values [16]. Although this number will be higher when aiming to distinguish European and Middle East/North African ancestries, applying likelihood thresholds is a suitable way to minimize ancestry inference error in forensic practice.

Acknowledgements

This research is part of the European Forensic Genetics Network of Excellence, funded by the European Union Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 285487. AFA is supported by a post-doctorate grant funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (Modalidade B, ED481B 2018/010).

References

[1] L. Cavalli-Sforza, P. Menozzi, A. Piazza, The History and Geography of Human Genes, Princeton University Press, Princeton, NJ, US, 1995.

[2] V. Grugni, V. Battaglia, B. Hooshiar Kashani, S. Parolo, N. Al-Zahery, A. Achilli, A. Olivieri, F. Gandini, M. Houshmand, M.H. Sanati, A. Torroni, O. Semino, Ancient migratory events in the Middle East: New Clues from the Y-chromosome variation of modern Iranians, PLoS One. 7 (2012).

[3] D.A. Badro, B. Douaihy, M. Haber, S.C. Youhanna, A. Salloum, M. Ghassibe-Sabbagh, B. Johnsrud, G. Khazen, E. Matisoo-Smith, D.F. Soria-Hernanz, R.S. Wells, C. Tyler-Smith, D.E. Platt, P.A. Zalloua, Y-Chromosome and mtDNA Genetics Reveal Significant Contrasts in Affinities of Modern Middle Eastern Populations with European and African Populations, PLoS One. 8 (2013)

[4] L. Quintana-Murci, R. Chaix, R.S. Wells, D.M. Behar, H. Sayar, R. Scozzari, C. Rengo, N. Al-Zahery, O. Semino, A.S. Santachiara-Benerecetti, et al., Where West Meets East: The Complex mtDNA Landscape of the Southwest and Central Asian Corridor, Am. J. Hum. Genet. 74 (2004) 827–845.

[5] M. Haber, D.E. Platt, D.A. Badro, Y. Xue, M. El-Sibai, M.A. Bonab, S.C. Youhanna, S. Saade, D.F. Soria-Hernanz, A. Royyuru, R.S. Wells, C. Tyler-Smith, P.A. Zalloua, Influences of history, geography, and religion on genetic structure: The Maronites in Lebanon, Eur. J. Hum. Genet. 19 (2011) 334–340.

[6] M. Haber, D. Gauguier, S. Youhanna, N. Patterson, P. Moorjani, L.R. Botigué, D.E. Platt, E. Matisoo-Smith, D.F. Soria-Hernanz, R.S. Wells, J. Bertranpetit, C. Tyler-Smith, D. Comas, P.A. Zalloua, Genome-Wide Diversity in the Levant Reveals Recent Structuring by Culture, PLoS Genet. 9 (2013).

[7] C. Phillips, A. Salas, J.J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Álvarez-Dios, M. Calaza, M.C. de Cal, D. Ballard, M.V. Lareu, Á. Carracedo, Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Sci. Int. Genet. 1 (2007) 273–280.

[8] R. Pereira, C. Phillips, N. Pinto, C. Santos, S.E.B. dos Santos, A. Amorim, Á. Carracedo, L. Gusmão, Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing, PLoS One. 7 (2012).

[9] C. Phillips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci. Int. Genet.

18 (2015) 49–65.

[10] J.M. Galanter, J.C. Fernandez-Lopez, C.R. Gignoux, J. Barnholtz-Sloan, C. Fernandez-Rozadilla, M. Via, A. Hidalgo-Miranda, A. V. Contreras, L.U. Figueroa, P. Raska, et al., Development of a panel of genome-wide ancestry informative markers to study admixture throughout the americas, PLoS Genet. 8 (2012).

[11] C. Phillips, W. Parson, B. Lundsberg, C. Santos, A. Freire-Aradas, M. Torres, M. Eduardoff, C. Børsting, P. Johansen, M. Fondevila, N. Morling, P. Schneider, Á. Carracedo, M. V. Lareu, Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set, Forensic Sci. Int. Genet. 11 (2014) 13–25.

[12] O. Bulbul, G. Filoglu, T. Zorlu, H. Altuncul, A. Freire-Aradas, J. Sochtig, Y. Ruiz, M. Klintschar, S.Triki Fendri, A. Rebai, C. Phillips, M.V. Lareu, A. Carracedo, P.M. Schneider, Inference of biogeographical ancestry across central regions of Eurasia, Int. J. Legal Med., 130 (2016), 73-79.

[13] L. Cherni, A.J. Pakstis, S. Boussetta, S. Elkamel, S. Frigi, H. Khodjet-El-Khil, A. Barton, E. Haigh, W.C. Speed, et al.,, Genetic variation in Tunisia in the context of human diversity worldwide, Amer. J. Physical Anthropol. 161 (2016) 62-71.

[14] O. Bulbul, L. Cherni, H. Khodjet-El-Khil, H. Rajeevan, K.K. Kidd, Evaluating a subset of ancestry informative SNPs for discriminating among Southwest Asian and circum-mediterranean populations, Forensic Sci. Int. Genet. 23 (2016) 153-158.

[15] O. Bulbul, W.C. Speed, C. Gurkan, U. Soundararajan, H. Rajeevana, A.J. Pakstis, K.K. Kidd, Improving ancestry distinctions among Southwest Asian populations, Forensic Sci. Int. Genet. 35, (2018) 14-20.

[16] C. Phillips, A. Freire-Aradas, A.K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F.S. Rech, M.D.P. Carceles, Á. Carracedo, P.M. Schneider, M. V. Lareu, Eurasiaplex: A forensic SNP assay for differentiating European and South Asian ancestries, Forensic Sci. Int. Genet. 7 (2013) 359–366.

[17] H.M. Cann, C. de Toma, L. Cazes, M.F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W.F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, et al., A human genome diversity cell line panel, Science 296 (2002) 261–262.

[18] N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard, Informativeness of Genetic Markers for Inference of Ancestry, Am. J. Hum. Genet. 73 (2003) 1402–1422.

[19] Excoffier L, Analysis of population subdivision, Handbook of statistical genetics, in: John

Wiley & Sons, Chichester, UK, 2001.

[20] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-sforza, R.M. Myers, Worldwide human relationships inferred from genome-wide patterns of variation, Science 319 (2008) 1100–1104.

[21] J. Amigo, A. Salas, C. Phillips, Á. Carracedo, SPSmart: Adapting population based SNP genotype databases for fast and comprehensive web access, BMC Bioinformatics. 9 (2008) 1–6.

[22] J. Costas, A. Salas, C. Phillips, Á. Carracedo, Human genome-wide screen of haplotype-like blocks of reduced diversity, Gene. 349 (2005) 219–225.

[23] P. Johansen, J.D. Andersen, C. Børsting, N. Morling, Evaluation of the iPLEX® Sample ID Plus Panel designed for the Sequenom MassARRAY® system. A SNP typing assay developed for human identification and sample tracking based on the SNP*for*ID panel, Forensic Sci. Int. Genet. 7 (2013) 482–487.

[24] L. Excoffier, H.E.L. Lischer, Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (2010) 564–567.

[25] Bonferroni, C. E., Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 1936.

[26] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics. 155 (2000) 945–959.

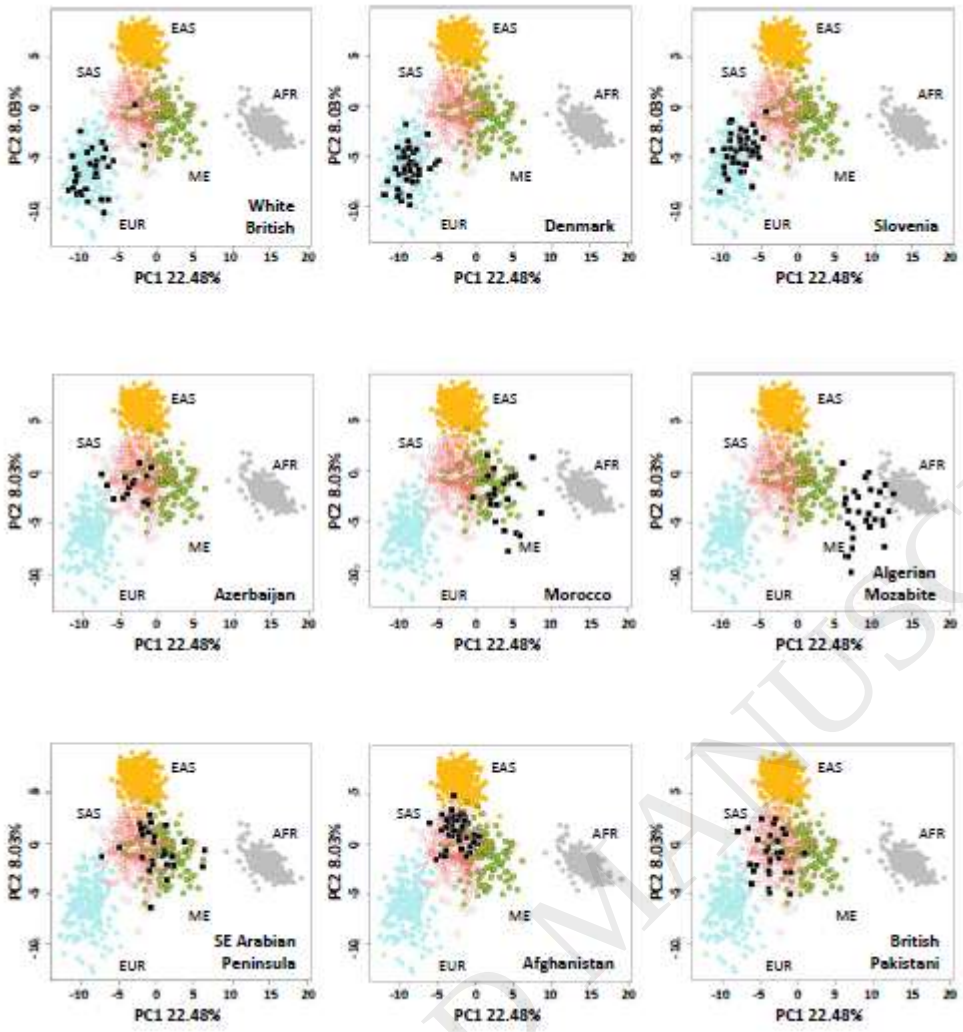
[27] D. Falush, M. Stephens, J.K. Pritchard, Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies, Genetics. 164 (2003) 1567–1587.

[28] N. Kopelman, J. Mayzel, M. Jakobsson, N. Rosenberg, I. Mayrose, Clumpak: a program for identifying clustering modes and packaging population structure inferences across K, Mol. Ecol. Resour. 15 (2015) 1179–1191.

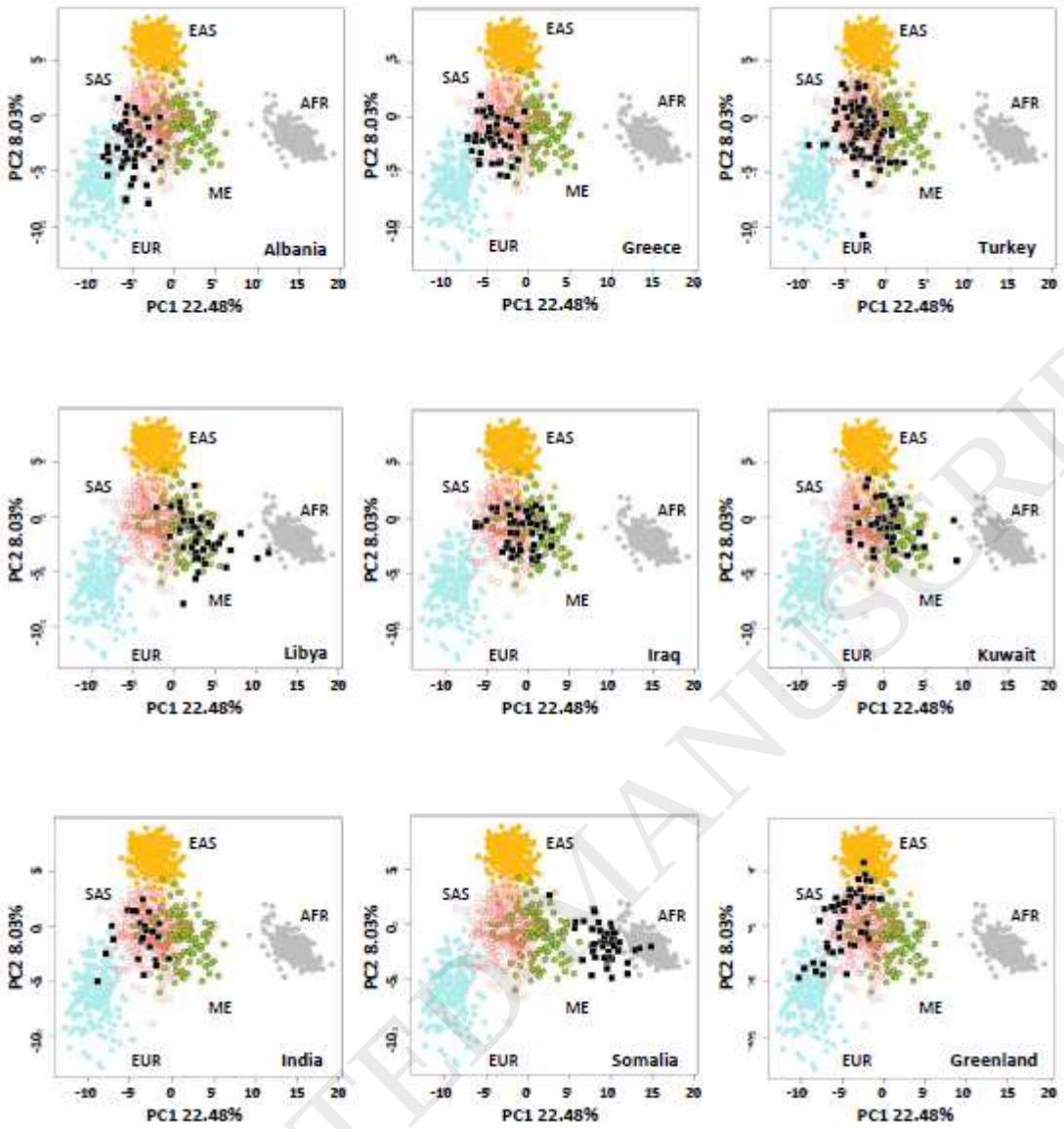
[29] M. Kayser, P. de Knijff, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (2011) 179–192.

Figure legends

Fig. 1. Map of the Middle Eastern and North African regions (with country borders shown for orientation).



A



B

Fig. 2A/B. Bi-dimensional principal component analyses of nineteen study populations following the west to east European to South Asian cline direction, plus additional populations of Somalia and Greenland. Five population groups: African (AFR, gray), European (EUR, blue), Middle East (ME, green), South Asian (SAS, red) and East Asian (EAS, orange) were used as reference clusters.

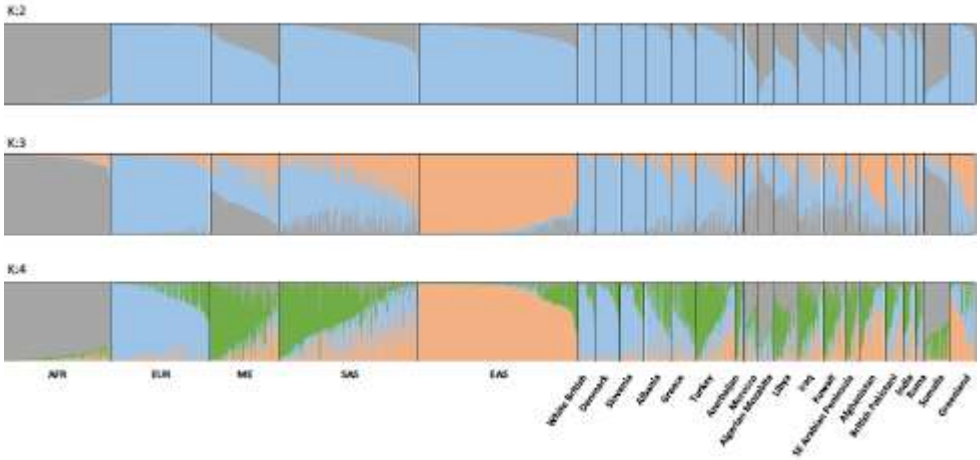


Fig. 3. STRUCTURE plots from K=2 to K=4 using 111 SNPs (NAME panel) for the nineteen study populations assessed. Genotypes for the samples used as reference groups: African (AFR); European (EUR); South Asian (SAS); East Asian (EAS) samples were obtained from 1000 genomes, and Middle East (ME) from the CEPH panel.

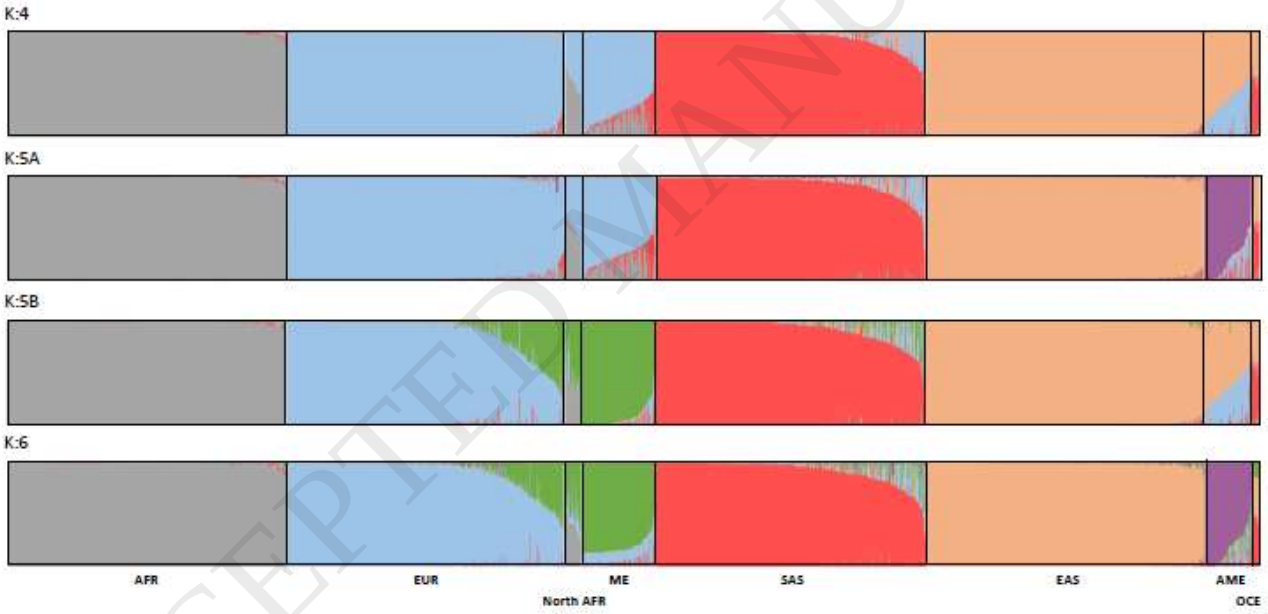


Fig. 4. STRUCTURE plots from K=4 to K=6 using 237 SNPs (NAME panel plus Global AIMs) for the seven major population groups (as shown in Fig. 3) plus North Africa from the CEPH panel. Two alternative clustering patterns are shown at K=5; identifying an American fifth genetic cluster (K:5A); or a Middle East fifth genetic cluster (K:5B) from the major and minor modes of the CLUMPAK analysis.